Building a land data assimilation community to tackle technical challenges in quantifying and reducing uncertainty in land model predictions

Natasha MacBean[a], Hannah Liddy[b,c], Tristan Quaife[d], Jana Kolassa[e,f], Andrew Fox[g]

[a] *Department of Geography, Indiana University, Bloomington, IN 47405, USA.*

[b] *Center for Climate Systems Research, Earth Institute, Columbia University, New York, NY 10025, USA.*

[c] *NASA Goddard Institute for Space Studies, New York, NY 10025, USA.*

[d] *National Centre for Earth Observation, University of Reading, RG6 6BB, UK.*

[e] *Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA.*

[f] *Science Systems and Applications, Inc., Lanham, MD 20706, USA.*

[g] *Joint Center for Satellite Data Assimilation, UCAR, Boulder, CO 80301, USA.*

Corresponding author: Natasha MacBean, nlmacbean@gmail.com

Information Box

AIMES Land Data Assimilation Working Group Virtual Workshop on "Tackling Technical Challenges in Land Data Assimilation"

What: Over 100 participants from the earth system modeling and numerical weather prediction land data assimilation communities met virtually to discuss technical challenges faced in developing land data assimilations systems, possible solutions and a roadmap for addressing those challenges, and ideas for building a land DA community to facilitate future collaborations and knowledge exchange.

Where: Virtual

When: 14th - 16th June, 2021

Website: https://aimesproject.org/lda_workshop/

Why is land data assimilation (DA) necessary?

The regional to global scale process-based land models that form part of numerical weather prediction (NWP) systems or earth system models (ESMs) have rapidly increased in their complexity over the past few decades (Prentice et al., 2015; Fisher and Koven, 2020). In parallel, there is a growing wealth of terrestrial observations that can be used to confront land models, including long-running observation/experiment campaigns (e.g., NSF LTER sites and DOE NGEE-Tropics and Arctic) and satellite missions or products (e.g., Landsat and MODIS from NASA, ESA CCI Soil Moisture), the synthesis of site-based and experimental manipulation data into networks (e.g., FLUXNET, ICOS, the International Soil Moisture Network, SAPFLUXNET, Drought-Net, FACE), novel ground-based observations (e.g., tree ring data, carbonyl sulphide, radiocarbon measurements), and a new array of space-based observations of global carbon, water, and energy cycles (e.g., hyperspectral and lidar instruments such as GEDI and HUSUI on the ISS, and solar induced fluorescence from platforms like OCO-2 and TROPOMI) at higher spatial, temporal, and spectral resolutions than ever before. Despite the increasing complexity of models and density of land-based data, uncertainty in land model projections remains high.

While there has been a concerted effort to use terrestrial observations for model evaluation; the number of studies that use these data for quantifying and reducing uncertainty in land model parameters and states via a statistical data assimilation (DA) framework is small in comparison. This is primarily due to the computational expense and technical challenges associated with implementing such a global-scale land DA system. However, land models urgently need to be confronted with a wide range of data to optimize model parameters, initialize surface states, and to address model structural uncertainty. Without such efforts, we cannot quantify or reduce uncertainty associated with individual model projections, and the inter-model spread in weather forecasts and predictions of land-atmospheric interactions or carbon-climate feedbacks will remain high (Arora et al., 2020).

The challenge of developing land DA systems

A number of land modeling groups spanning different modeling communities (carbon, hydrology, land surface/earth system modeling (LSM/ESM) and numerical weather prediction (NWP)) have devoted significant resources into developing global-scale land DA systems. However, this technical development work does not get the level of exposure in the literature or in conference talks commensurate with the resources needed to complete that work because publications and presentations are naturally focused on scientific questions. For the same reason, development of DA systems is typically not the focus of grant proposals nor calls for proposals by funding agencies. Nonetheless, their development can consume much of the time on a given research project due to the considerable complexities and technical challenges faced when implementing DA methods within land models. As a result, this largely hidden work and advancement of "best practices" or solutions for technical challenges are not widely publicized or discussed at scientific meetings. The community is aware that there is a wealth of experience and knowledge within each land DA group, but there needs to be an easier framework for collaboration and knowledge exchange between land DA groups so that all groups can more readily benefit from existing research. Such a community-sharing framework might inspire new avenues of research to help solve these challenges and progress our land DA systems more efficiently.

Starting to build a land DA community

In 2020, the Analysis, Integration and Modeling of the Earth System (AIMES) global research project of Future Earth formed a Land Data Assimilation Working Group (WG)

(Schimel et al., 2015; https://aimesproject.org). The primary objective of the AIMES Land DA WG is to build a community of researchers working on integrating observations with models of terrestrial processes using cutting-edge mathematical techniques. Our secondary objective is to publicize the use of these techniques across the land modeling communities so uncertainty quantification and reduction becomes routine practice. Bringing the implementation of land DA methods to the forefront holds the potential to dramatically improve our understanding and quantification of uncertainties in land model predictions.

Kick-starting a land DA community with a virtual workshop

To begin the process of establishing a land DA community and facilitating knowledge exchange between different land DA groups, we held a 3-half day virtual workshop in June 2021 with over 100 attendees participating each day. The workshop highlighted a range of DA methods used within the community, addressed the challenges facing different modeling groups, and identified strategies for addressing these challenges. We defined three broad themes for each day: Day 1) Applicability of data assimilation approaches across different land modeling groups; 2) Emerging techniques in land DA; and Day 3) Challenges in dealing with observations. Cross-cutting themes that spanned multiple days of the workshop included addressing issues related to error characterization and the different spatial and temporal scales over which we assimilate data. Each day, six invited speakers gave 15-minute presentations. Following the talks, breakout groups discussed targeted questions relevant to the theme of each day as well as the following: i) What factors are hindering progress in this area and what are the next steps to address these challenges?; and ii) What do you think the land DA WG and/or a land DA community could do to facilitate progress in this area? The plenary session featured brief summaries of each breakout group discussion from the lead moderator of each group as summarized here.

What are the technical challenges we face in land DA?

Challenges faced by both NWP and ESM communities

Bringing together the NWP and ESM DA communities provided the opportunity to learn more about each other's objectives, technical issues, and solutions being developed to improve land model predictions. In terms of objectives, the NWP community is under operational pressure to improve forecast skill with a focus on updating model states with DA, whereas the ESM land community is under pressure to improve and reduce uncertainty in predictions of longer timescale variables and has therefore typically had more of a focus on constraining model parameters. The identified technical issues and some possible solutions included:

- Ensemble DA methods are used by both the NWP and ESM communities; however, many questions remain about the best approaches for initializing those ensembles, such as how to maintain appropriate ensemble spread (i.e. in such a way as it represents errors in the system appropriately) and how to apply techniques such as localization to avoid the impact of spurious correlations in space and/or time?
- Defining observation and model structural uncertainty is hard and approaches for doing so are still somewhat "ad-hoc" and generally do not account for error covariances. Possible approaches for tackling this issue were discussed with an acknowledgement that the community also needs to communicate our observation uncertainty needs to data providers. The land DA community could learn more from methods used in the atmosphere DA community, who have worked on this problem in more detail because of the greater number of observations assimilated. A collaborative study focusing on this topic would help to advance DA techniques in this area.
- To avoid systematic errors in NWP and ESMs, there is a need to either observe and assimilate observations more directly linked to what is represented in the models, perform bias corrections, develop models so all relevant processes are accounted for and/or use machine learning (ML) to bridge the gap between process-based models and observations. These issues are linked to the scale of prediction and observation.

Keeping the DA system up-to-date with the land model is difficult given the rapid pace of model developments. The community may need to adopt DA techniques that are more flexible in terms of coupling to the land model and/or adapt land models to better facilitate their interaction with DA systems. Computational efficiency of the algorithms is paramount for both communities. More specific funding for longer-term technical developments and computational infrastructure building and maintenance, such as the inter-agency Joint Effort for Data Assimilation (www.jcsda.org/jcsda-project-jedi) in the US, would be enormously helpful.

DA in Modeling of Human Activity

To date, DA has rarely been used to improve land model predictions of human activity. Therefore, discussion focused on how we can facilitate the use of DA in crop and land management models. Key questions remain as to the best use of DA in constraining models of human activity: When accounting for agriculture land management (including fire suppression) in models, is it better to implement a process-based model and use DA to calibrate parameters and initialize surface states? Or, do we use state DA in place of process-based models where we do not have enough information on the processes and/or the processes vary too much over space and time to be worth representing in models? Other uses of DA in this area could be to accurately account for historical land use change on carbon stocks. Research in this area is in its relative infancy and the answers to these questions may depend on the specific type of human activity or process being considered. More dedicated discussion and collaboration to make use of DA for modeling human activity is planned and will be beneficial to all land DA communities.

Optimal Model Complexity

Model complexity was identified as a limiting factor when using DA to constrain uncertainty in models, and it is unclear if process-based model complexity has vastly outpaced observing system "dimensionality" (i.e., volume and/or diversity of land surface measurements), or if we have just not yet developed DA systems that can make full use of the information content of all available data. In general, many workshop participants felt that process-based global land models should continue to include as much relevant process representation as possible, although they noted that the level of complexity depends on the objectives or research questions being asked, the scale of the analysis, and on the available data. Simpler models are easier to confront with data, but simple model skill may be demonstrably biased if it does not include the right processes. However, confronting simpler models with data might be a cost-effective solution for testing technological DA developments while waiting for the computational advances needed to perform DA with more complex models.

Novel Observations

There is broad agreement in the land DA community of the potential benefit of including as many novel and traditional datasets as possible to constrain models, although there needs to be greater use of rigorous methods (e.g., sensitivity analyses) for how to select the most important datasets needed to inform our models. Based on our collective experience in this area, two main issues arose when considering how to make the best use of available data: 1) Constructing the correct "observation operator" (i.e., the additional model that links the relevant land model variable to the data) is often complicated by a lack of process understanding and/or the need to constrain additional parameters, different spatial and temporal resolutions, and localization and representativity issues; and 2) Acquiring enough information on data characteristics and uncertainties is difficult when this information is under-reported. One solution when assimilating satellite data might be to assimilate "low-level" satellite products such as reflectances or brightness temperatures that are closer to the variables directly measured by a satellite instrument. Assimilating raw data is preferable because modelers do not have to account for uncertainties in the retrieval algorithm used to derive higher level products; instead, the modelers themselves can implement the retrieval algorithm (or observation operator) within their land model. This allows the modeler to have full knowledge of uncertainties associated with that observation operator and to ensure its design is compatible with the processes represented in their land model. Where "higher level" products derived from the raw data via a satellite data retrieval algorithm are assimilated, it is important to have detailed uncertainty information and metadata about those data products. The NWP community faces an additional challenge of needing to assimilate data operationally in real-time, which requires development of efficient pre-processing frameworks and advances in data sharing agreements and infrastructure. Wherever possible, model developments should be guided by the available data and actually represent what is measured. Future efforts in this area will bring together the land DA community with data providers to learn more about each other's needs.

Machine Learning in DA

Machine learning (ML) applications are still nascent in land DA but investing in developing and testing these techniques has the potential to further maximize the use of data in a number of ways: For instance, process-based observation operators could be replaced with ML approaches where insufficient knowledge of the processes involved is an issue. However, there is very little effort in that direction, largely because of limited understanding of ML techniques as well as a healthy amount of skepticism towards them. It is possible that neither the process-based observation operators or the ML approaches can solely address the challenges we are facing in linking models to data; instead, there might be benefit in moving towards hybrid approaches that combine elements of both. Other possible uses of ML techniques in land DA include improving model benchmarking and identifying the most important processes we need to constrain with data. A proposed solution to advance research in this area was fostering collaboration between the process-based modelling and ML communities, possibly in the form of a hackathon, or a DA model intercomparison project to compare traditional and ML DA techniques as well as hybrid techniques. The goal being to enhance our understanding of the capabilities and limitations of ML algorithms, to move to a more widespread adaptation of these techniques, and to consider the generation of hybrid approaches.

Building a land DA community - more communication needed!

This workshop brought together a broad range of land surface modelers working on different land data assimilation challenges. The benefit of seeing talks focused on the technical challenges from these different communities was widely recognized as useful and important for fostering communication across different communities, with the objective of being more cognizant of each community's requirements, limitations, tools, and challenges. The following communication was identified as necessary to ultimately be more 'useful' to each other:

- NWP and ESM land DA communities to share experience of how to solve technical and computational challenges;
- Land DA practitioners and the crop and land management modeling community to discuss best ways to move forward with using land DA to improve modeling of impacts of human activities;
- Land DA and land model development communities to discuss: (1) what metrics should be used to evaluate land DA efforts, other than the impact on the atmospheric forecast or climate projection skill; and (2) how to continue developing models in a way that facilitates the assimilation of (novel) land observations;
- Data producers and the land DA community to (1) make the land DA community aware of novel datasets that could be useful to them and approaches used for deriving "higher level" products and (2) make the data producers aware of the data requirements and information on data characteristics needed to ensure the data are

properly used in assimilation experiments (e.g., spatial/temporal resolution, record continuity, the types of variables needed, detailed information on observation uncertainties);

- Land DA community to discuss more with the ML community to develop new approaches for tackling some technical land DA challenges.

Next steps in building a land DA community

Through this workshop, we achieved (1) improved knowledge of the range of data assimilation methods used by different land modeling groups, (2) improved understanding of the challenges facing different modeling groups and potential strategies for addressing those challenges, and (3) the first step toward establishing a land DA community to increase collaboration on tackling these technical challenges and promoting the routine use of data assimilation methods within the wider land and earth system modeling community.

To build on the momentum gained from the workshop and the positive responses to our goal of building a community, the AIMES Land DA WG identified more in-depth, topic-specific activities to spur future events. These events will take place virtually for the foreseeable future; however, we are planning in-person social events at conferences to facilitate the personal side to building the community. In the near-term, we also hope the community will facilitate collaborative organization of conference sessions and focused workshops (or a series of shorter meetings) to address best approaches for using ensemble methods, how to use DA in crop and land management modeling, and continuing discussion on the similarities and differences between NWP and ESM land DA communities. We are already in the early stages of planning a data provider half-day workshop to learn more about novel datasets and observation networks.

In addition to meetings and workshops, we aim to facilitate community collaborative studies on global sensitivity analyses, a calibrated model intercomparison project to illuminate model structural uncertainty, and how best to define observation error covariances. We would like to expand on the significant early career involvement during the workshop to create an early career researcher wing of the land DA community. We will solicit community involvement in sharing educational materials and training events. These resources – and all information related to community events – will be collated and published on a planned Land DA Community website. In the long-run, our vision is that the Land DA Community will take on a life of its own, continuing our common goal of knowledge sharing and collaboration on synergistic activities to overcome land data assimilation related challenges.

To get involved, sign up for the Land DA Community listserv found at https://aimesproject.org/lda_workshop/ to facilitate collaboration and knowledge sharing. In the coming months and years, the AIMES Land DA WG, in close collaboration with the community, will help to write a land DA review and agenda-setting article that outline the state of the science and community needs to advance land DA research. We hope this will help us to promote the need for land model uncertainty quantification and advocate for longer-term funding for developing and maintaining land DA systems.

Acknowledgments.

This article is an outcome of the workshop ‘Tackling Technical Challenges in Land Data Assimilation’ held virtually in June 2021. The workshop was organized by the Land DA Working Group of the Analysis, Integration, and Modeling of the Earth System (AIMES) global research project of Future Earth. We thank the invited speakers for their fantastic talks: Patricia De Rosnay (ECMWF), Eunjee Lee (NASA GSFC), Bertrand Bonan (CNRM), Marko Scholze (Lund University), Breo Gomez (UK Met Office), Sujay Kumar (NASA GSFC), Jianzhi Dong (MIT), Ewan Pinnington (University of Reading), Istem Fer (Finnish Meteorological Institute), Joanne Waller (UK Met Office), Moha El Gharamti (NCAR/UCAR), Anthony Bloom (NASA JPL), Nina Raoult (LSCE), Susan Steele-Dunn (Delft University of Technology), Jina Jeong (Vrije Universiteit Amsterdam), Ann Raiho (NASA GSFC/University of Maryland), Clara Draper (NOAA), and Manuela Girotto (UC Berkeley). Recorded presentations are available for viewing on the AIMES YouTube Channel: https://www.youtube.com/playlist?list=PLXPEVXGzl0mhX10wtKNEflaZ-foyT-767. We thank Clara Draper (NOAA), Min Huang (GMU), Alex Ruane (NASA GISS), Clément Albergel (ESA Climate Office), Forrest Hoffman (ORNL), José Gómez-Dans (University College London), Marko Scholze (Lund University), Nina Raoult (LSCE), Anthony Bloom (NASA JPL), Kashif Mahmud (Indiana University), Joanne Waller (UK Met Office), Philippe Peylin (LSCE), Manuela Girotto (UC Berekeley), Andria Dawson (Mount Royal University), Mike Dietze (Boston University), Adrian Rocha (University of Notre Dame), Harrie-Jan Hendricks-Franssen (Forschungszentrum Julich GmbH), and Sujay Kumar (NASA GSFC) for leading the breakout group discussions during the workshop.

REFERENCES

Arora, V. K., and Coauthors, 2020: Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models. *Biogeosciences*, **17**, 4173–4222, https://doi.org/10.5194/bg-17-4173-2020.

Fisher, R. A., and C. D. Koven, 2020: Perspectives on the Future of Land Surface Models and the Challenges of Representing Complex Terrestrial Systems. *J. Adv. Model. Earth Syst.*, **12**, https://doi.org/10.1029/2018MS001453.

Prentice, I. C., X. Liang, B.E. Medlyn, and Y. P. Wang, 2015: Reliable, robust and realistic: the three R's of next-generation land-surface modelling. *Atmospheric Chem. Phys.*, **15**, 5987–6005, https://doi.org/10.5194/acp-15-5987-2015.

Schimel, D., K. Hibbard, D. Costa, P. Cox, and S. van der Leeuw, 2015: Analysis, Integration and Modeling of the Earth System (AIMES): Advancing the post-disciplinary understanding of coupled human–environment dynamics in the Anthropocene. *Anthropocene*, **12**, 99–106, https://doi.org/10.1016/j.ancene.2016.02.001.